# Iterative Self-Projection (ISP)

Iterative Self-Projection is a training procedure in which a neural network repeatedly relearns its own behavior rather than repeatedly relearning the original target function.

## Procedure

1. Train a neural network on a standard dataset of input-target pairs.
   Input: x
   Target: y
2. After training, run all training inputs through the network and record the outputs.
   Input: x
   Network Output: $y_1$
3. Construct a new training dataset using the original inputs and the recorded outputs.
   Input: x
   Target: $y_1$
4. Retrain the network from scratch or from its current weights using this new dataset.
5. Repeat the process multiple times, producing successive generations:
   $y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow \ldots$
6. Periodically retrain against the original targets y to maintain grounding and prevent uncontrolled drift.

## Motivation

The original targets may contain noise, inconsistencies, or details that are difficult for the architecture to represent. Each generation projects the learned function back onto the set of functions naturally expressible by the network.

The process may therefore:

- Smooth noisy behavior.
- Simplify decision boundaries.
- Reveal stable internal representations.
- Expose fixed points of the learning dynamics.

Alternatively, information may be lost over generations, leading to excessive simplification or representational collapse.

## Two-Network Variant

Network A undergoes Iterative Self-Projection and is allowed to drift.

Network B is trained to map the outputs or latent representations of A back to the original target values.

In this configuration:

- Network A develops an evolving internal representation.
- Network B maintains task grounding.
- The evolution of representation can be studied independently from task performance.

# Measurements

Useful quantities to track include:

- Training loss.
- Generalization error.
- Output entropy.
- Representation dimensionality.
- Information retained relative to the original targets.
- Convergence toward stable functions or attractors.